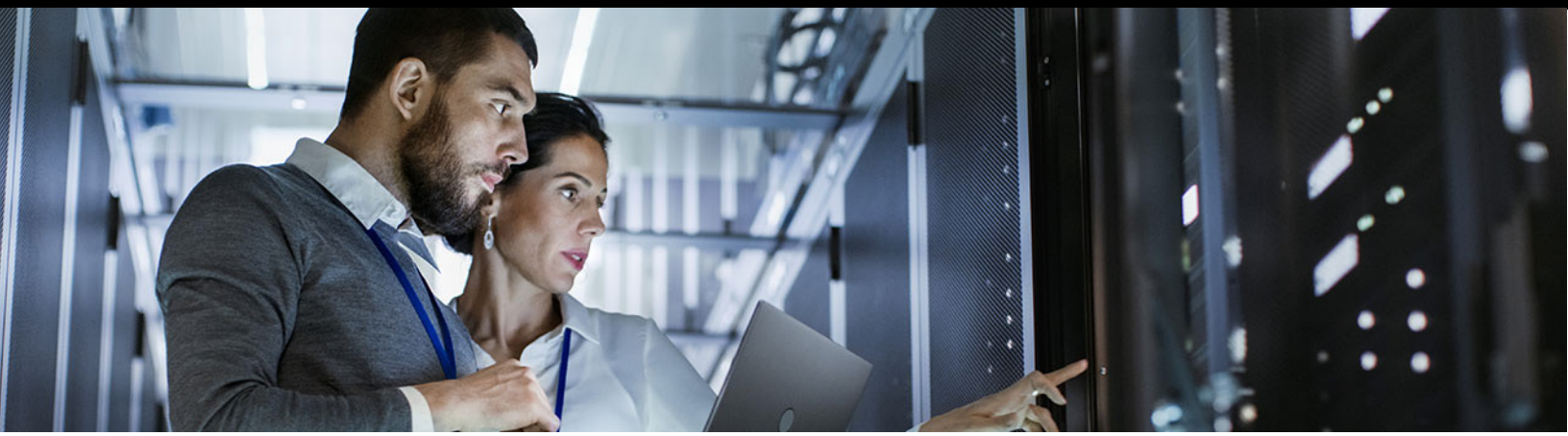

Micron® 6500 ION NVMe™ SSD Enables Better Ceph® Storage Performance and Better Resiliency Than Competitor's QLC SSD[1]

With the recent introduction of high-capacity NVMe SSDs and modern software, object storage solutions now offer more than enough performance for many demanding data center applications as all-flash configurations increasingly become the norm.

One popular solution for object storage is Ceph Storage Community Edition. It is a scalable, simple, open storage software package for modern data-centric applications – from artificial intelligence and machine learning to data analytics and emerging cloud solutions.[2] When deploying Ceph, a major question now centers on which type of value-focused high-capacity NVMe SSD to use.

This technical brief shows how the 30.72TB Micron® 6500 ION SSD enables better cluster performance, more efficient CPU utilization (i.e., less time for the CPU to wait on storage), and better cluster resiliency (ability to quickly recover from failure) compared to the 30.72TB Solidigm® D5-P5316 (a QLC SSD).[3]

Test results clearly show that the Micron 6500 ION is an ideal fit offering high performance and massive capacity in the same object store.

Fast Facts

The Micron 6500 ION, a high-capacity NVMe SSD, improves scalability through its extreme capacity and high performance in Ceph object stores.

The 30.72TB Micron 6500 ION SSD test results show meaningful performance improvements in all tested workloads and a marked improvement in cluster durability compared to the 30.72TB Solidigm D5-P5316 SSD results.

3.5X **Peak improvement in sequential write performance**

49% **Peak improvement in random read performance**

62% **Peak improvement in mixed read/write performance**

31% **Better cluster resiliency (faster recovery time)**

micron.com/6500ION

1. Performance means throughput (GB/s). Resiliency means the time needed for the Ceph cluster to rebuild to its default data protection level after an SSD failure. Comparative statements refer to test results shown herein versus the Solidigm D5-P5316, a competing, Quad-Level Cell (4 bits per cell) NVMe, NAND-based SSD.
2. See https://www.redhat.com/en/blog/ceph-open-source-community-powering-red-hats-data-services-portfolio for additional information on Ceph Storage Community Edition.
3. Unformatted. 1 GB = 1 billion bytes. Formatted capacity is less.

Workload testing

Throughput performance is evaluated by executing multiple tests using various scaling parameters in RADOS bench[4] (this tool is provided as part of the Ceph package). This benchmark reports throughput performance in GB/s and represents the best-case object performance. Object I/O uses a RADOS gateway service operating on each load generation server (the configuration of RADOS gateway is beyond the scope of this document). We chose 256KB objects to reflect a balance between larger objects (which tend to produce higher throughput) and smaller objects (which tend to product lower throughput due to IO overhead). Additional information on object sizing is available here: https://docs.ceph.com/en/quincy/radosgw/config-ref/.

Performance results are represented on the y-axis (higher is better) while scaling, threads per instance, is shown on the x-axis (increasing from left to right). The Micron 6500 ION SSD results are shown in blue and the Solidigm D5-P5316 results are shown in grey. The maximum performance improvement between the two configurations is reflected in bold font. All tested performance values are shown for completeness.

- **Read tests** (100% sequential and 100% random): These tests use 60 instances of RADOS bench and scales the thread count per instance from 2 to 32.
- **Write tests** (100% sequential): These test use 16 threads per RADOS bench[4] instance and number of instances scales from 10 to 60. The number of instances is abbreviated "#instances" on the horizontal axis in the figures below.
- **Mixed IO tests**: These tests are a mix of read and write IO. They simultaneously use 60 instances of 100% read and 60 instances of 100% write for each test. Each instance scales from 2 to 32 threads.

Workload results

100% sequential read

Figure 1 represents 100% sequential read test results. The Micron 6500 ION SSD shows superior results for all tested threads per instance.

The Micron 6500 ION SSD **maximum performance improvement of 47%** occurs when the threads per instance = 16. The lowest performance improvement of 30% occurs when threads per instance = 4.

Performance improvements are similar for other tested threads per instance values.

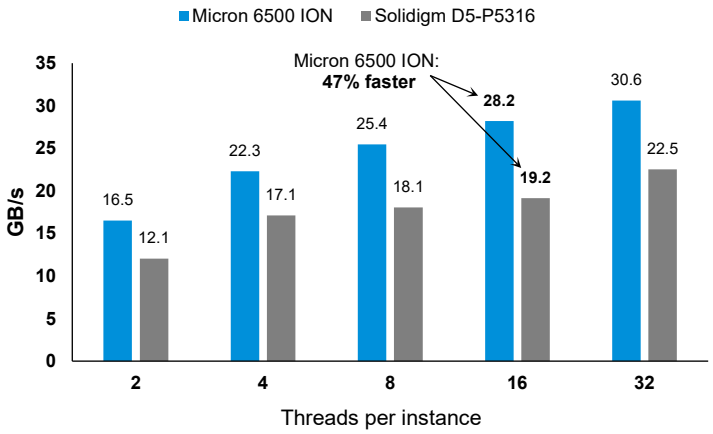

Figure 1: 100% Sequential read performance

100% sequential write

Figure 2 represents 100% sequential write test results. The Micron 6500 ION SSD shows superior results for all tested #instances.

The Micron 6500 ION SSD performance improvement reaches a **maximum improvement of 3.5X** when #instances = 10 and a minimum improvement of 6% when #instances = 60.

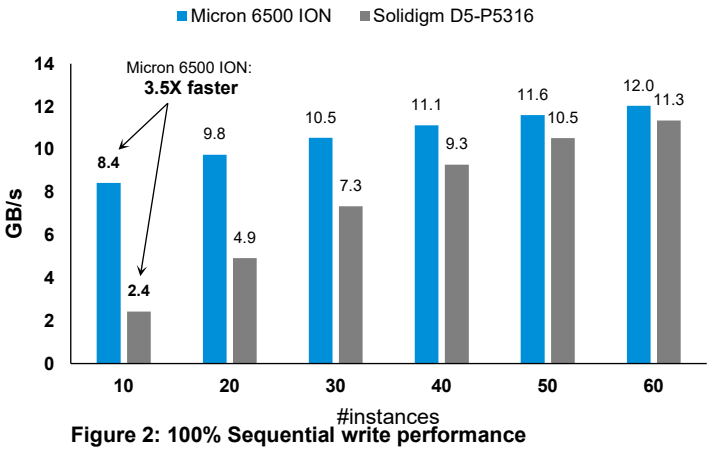

Figure 2: 100% Sequential write performance

4. Additional details on RADOS bench are available here: https://docs.ceph.com/en/latest/man/8/rados/#bench-options. Details on configuration a RADOS gateway are available here: https://docs.ceph.com/en/quincy/radosgw/config-ref/

100% random read

Figure 3 represents 100% random read test results. As seen in prior workloads, the Micron 6500 ION SSD shows superior results for all tested threads per instance.

The **maximum performance improvement of 49%** is seen with 16 threads per instance and the lowest improvement of 30% is seen at 2 threads per instance.

.

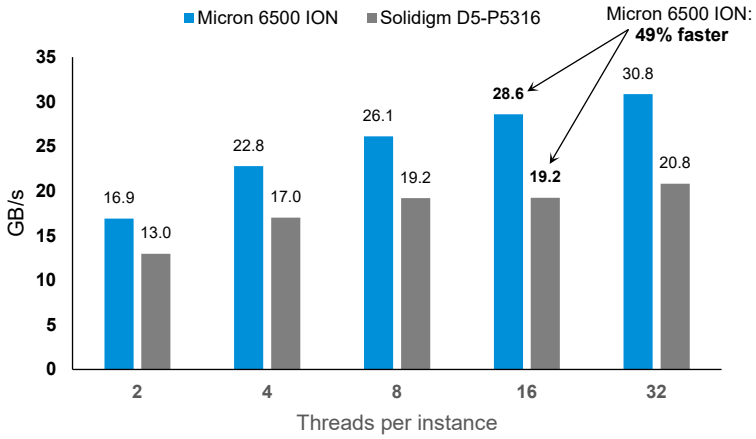

Figure 3: 100% Random read performance

Random read and sequential write

This mixed IO workload simultaneously uses 60 instances of 100% random read and 60 instances of 100% sequential write. Each instance scales from 2 to 32 threads per instance.

The difference in **maximum performance improvement of 27%** is seen with 32 threads per instance and the lowest improvement of 15% is seen at 8 threads per instance.

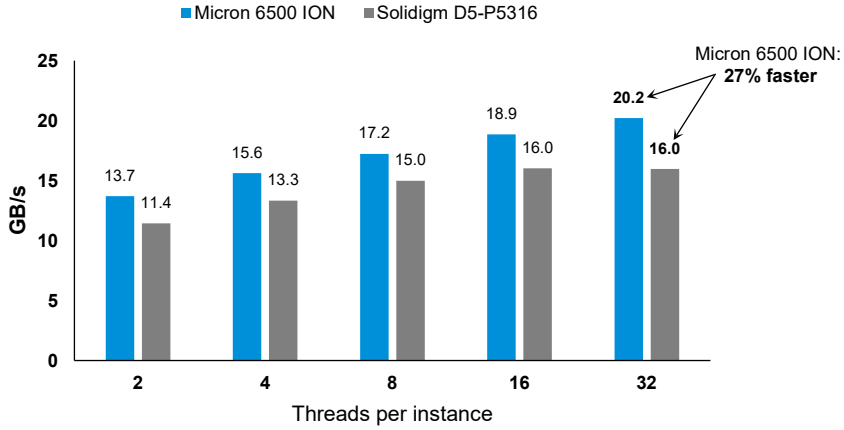

Figure 4: 100% Random read + 100% sequential write performance

Sequential read and sequential write

This mixed IO workload simultaneously uses 60 instances of 100% sequential read and 60 instances of 100% sequential write. Each instance scales from 2 to 32 threads per instance.

The **maximum performance improvement of 62%** is seen with 4 threads per instance and lowest improvement of 49% at 2 threads per instance.

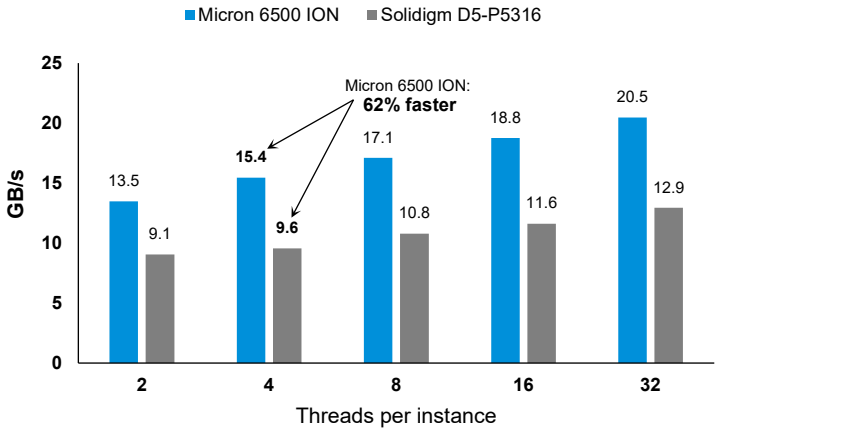

Figure 5: 100% Sequential read + 100% sequential write performance

CPU utilization testing

CPU utilization is evaluated by executing the same test workloads used to evaluate performance (again, using varying scaling parameters and the RADOS bench tool). The percentage of user CPU time (abbreviated "usr") is defined as the percentage of time the CPU uses to run application code. System CPU time (abbreviated "sys") is defined as the percentage of time the CPU uses to run the operating system (i.e., the kernel). The combination is abbreviated "CPU (usr+sys)." Higher values of CPU (usr+sys) indicate that more CPU time is spent doing productive tasks.

The CPU may also have to wait for storage to respond to an IO request. The percentage of time the CPU waits for IO to respond depends on the underlying storage system and is abbreviated "CPU IOWait." Lower CPU IOWait values are better because they indicate that the CPU is spending less time waiting on the storage device for an IO to finish.[5]

In the figures below, the CPU (usr+sys) and CPU IOWait values are represented on the y-axis while scaling is shown on the x-axis (increasing from left to right). The Micron 6500 ION SSD results are shown in blue and the Solidigm D5-P5316 results are shown in grey. Higher values of CPU (usr+sys) are better because the system's CPUs are doing productive work while lower values of CPU IOWait are better because those CPUs spend less time idle.

CPU utilization results

100% sequential read

Figures 6a and 6b represent 100% sequential read test CPU utilization.

Figure 6a shows that CPU utilization in the Micron 6500 ION configuration is higher than the for the Solidigm D5-P5316 SSD configuration. Figure 6b shows that the Micron 6500 ION storage configuration spends fewer cycles waiting for storage IO (lower IOWait values).

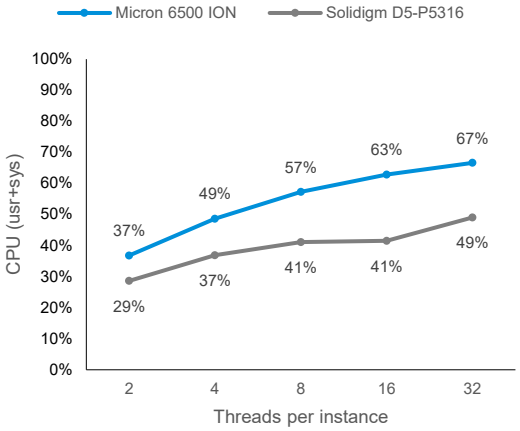

Figure 6a: 100% Sequential read CPU utilization

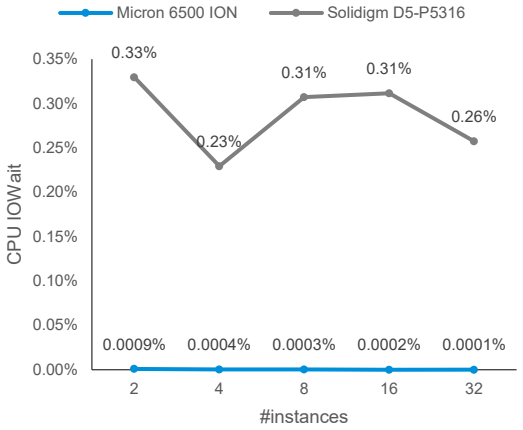

Figure 6b: 100% Sequential read CPU waiting on storage[6]

100% sequential write

Figures 7a and 7b represent 100% sequential write CPU utilization.

Micron 6500 ION configuration CPU (usr+sys) values are higher than the CPU (usr+sys) values for the Solidigm D5-P5316 configuration as seen in Figure 7a. The CPU IOWait values in Figure 7b show that the Micron 6500 ION is far more responsive; its CPU IOWait values are much lower across all tested #instances.

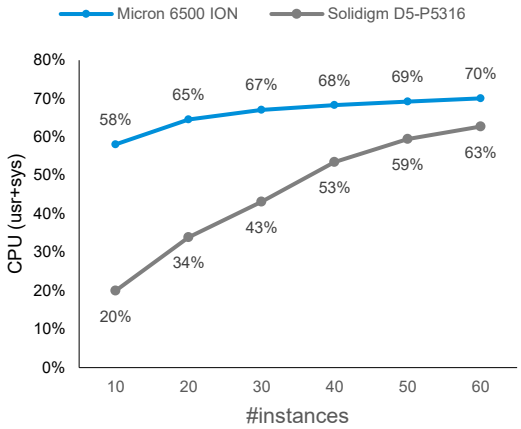

Figure 7a: 100% Sequential write CPU utilization

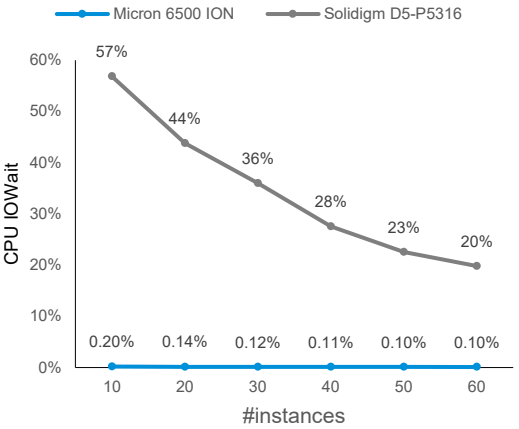

Figure 7b: 100% Sequential write CPU waiting on storage

5. See https://forums.oracle.com/ords/apexds/post/user-cpu-time-us-time-in-top-6003 for additional details.
6. Micron 6500 ION SSD y-axis values are shown to four decimal places in select figures. Showing fewer than four decimal places results in these values being displayed as 0.00%. The Solidigm D5-P5316 values are higher, needing only 2 decimal places to display.

100% random read

Figures 8a and 8b represent 100% random read test results.

As seen in prior workloads, the Micron 6500 ION SSD shows superior results for all tested threads per instance.

CPU IOWait values for the Micron 6500 ION SSD configuration so low that they are shown to 4 decimal places (instead of just 2) to help ensure visibility.

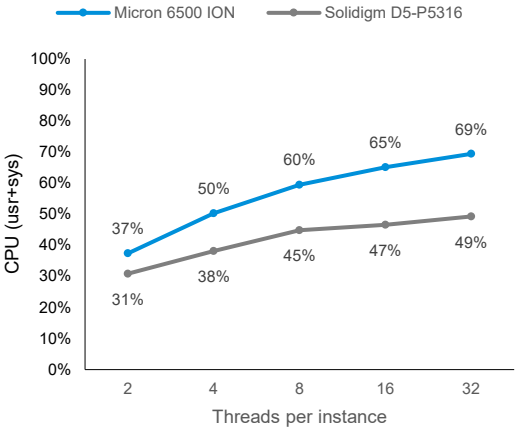

Figure 8a: 100% Random read CPU utilization

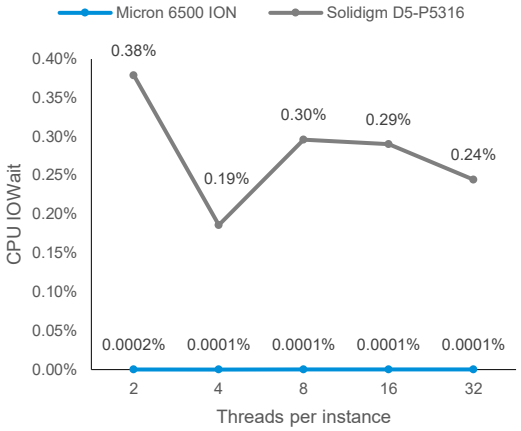

Figure 8b: 100% Random read CPU waiting on storage

Random read and sequential write

Figures 9a and 9b represent mixed IO workload results, which simultaneously use 60 instances of 100% random read combined with 60 instances of 100% sequential write.

Each instance scales from 2 to 32 threads per instance.

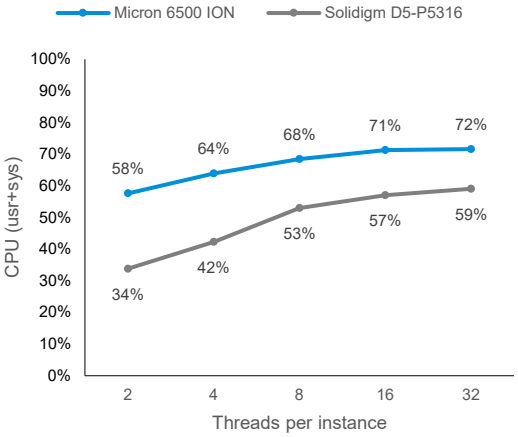

Figure 9a: Random read + sequential write CPU utilization

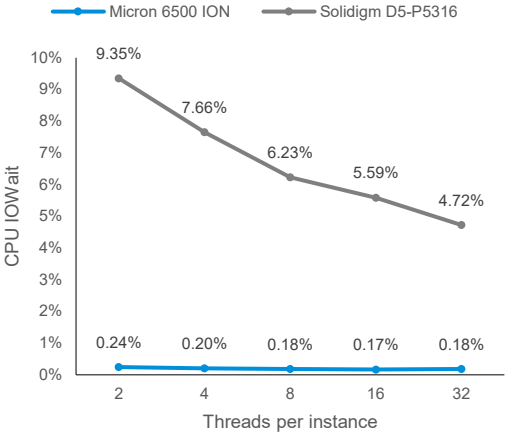

Figure 9b: Random read + sequential write CPU waiting on storage

Sequential read and sequential write

Figures 10a and 10b represent mixed IO workload results.

The workload simultaneously uses 60 instances of 100% sequential read with 60 instances of 100% sequential write.

Each instance scales from 2 to 32 threads per instance.

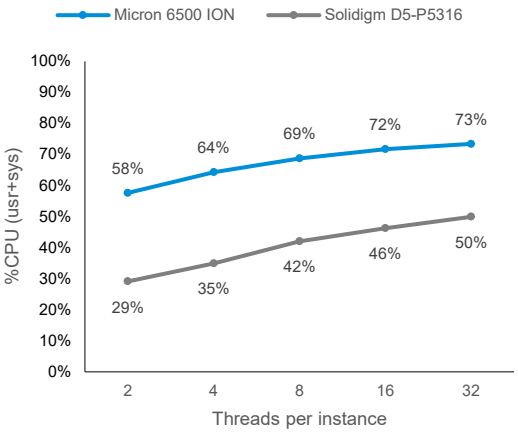

Figure 10a: Sequential read + sequential write CPU utilization

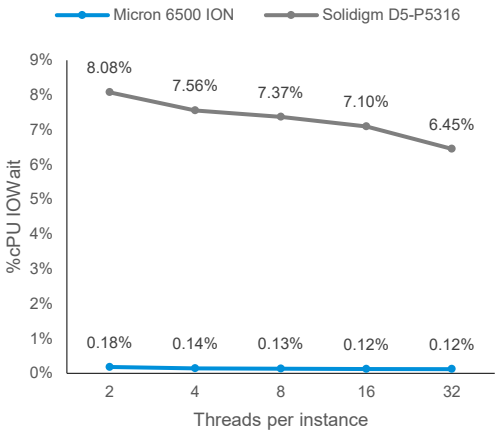

Figure 10b: Sequential read + sequential CPU waiting on storage

Resiliency testing

Cluster resiliency — the ability to recover from failure quickly and effectively — is important to data availability. For example, Ceph storage systems must recover from failure quickly while continuing to operate in order to minimize service disruption. This is especially true when a cluster node (and, hence, the cluster) is in a failed state (for example, when an SSD running the object storage daemon [OSD] fails).

Two tests are used to evaluate resiliency: 1) restore cluster health without an additional load on the cluster and 2) restore cluster health with an additional load (loaded test). The loaded test subjects the cluster to a continuous workload consisting of 60 read instances (16 threads, 256K objects) and 1 write instance (16 threads, 256K objects) during the rebuild process.

Micron's resiliency testing starts by filling the Ceph cluster to approximately 70% capacity (this is equal to about 700TB across all SSDs in these two different SSD configurations) and then verifying cluster health to ensure a consistent starting point. Four OSD service instances are stopped (each SSD has 4 OSDs), rendering that OSD node offline without the cluster rebalancing or rebuilding) to simulate a failure. Next, one SSD in the now offline node is removed and replaced with a secure-erased SSD. The failed OSD service is then redeployed (and its OSD instances started), and the timer starts.

The timer stops when the cluster returns to a healthy state (the Ceph storage degraded data indicator reads 0%, signaling that the rebuild process is complete). Note that while rebalancing may be occurring, the degraded data indicator value of 0% confirms that all data has been rebuilt and that cluster health has been restored.

Resiliency results

Figure 11a represents resiliency testing results with no additional load, while figure 11b represents these results with the noted load applied. The SSD tested is shown on the vertical axis of each figure while the cluster rebuild time (in hours) is shown on the horizontal axis.

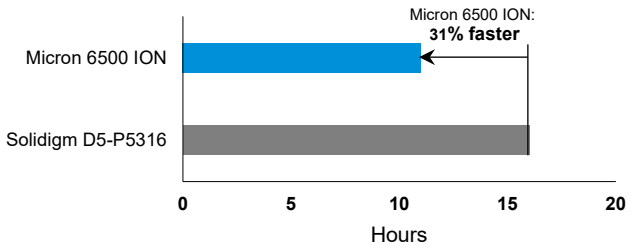

Figure 11a: Cluster rebuild time without load

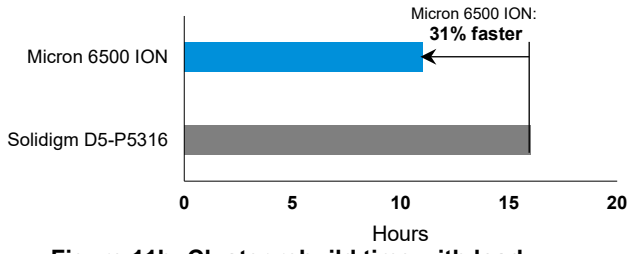

Figure 11b: Cluster rebuild time with load

Figures 12a and 12b indicate that the Micron 6500 ION cluster rebuilds in just **11 hours** (blue bar) while the Solidigm D5-P5316 cluster rebuilds take **16 hours** (grey bar) — whether a cluster is loaded or not. This shows a consistent **5-hour improvement**, or 31% advantage, in cluster resiliency (faster rebuild time) for the Micron 6500 ION.

Figure 12a represents cluster network throughput during loaded rebuild, while Figure 12b reflects CPU utilization during loaded rebuild. Both CPU (usr+sys) and CPU IOWait are shown for both clusters. The Micron 6500 ION **cluster network throughput is 13% higher** than the Solidigm D5-P5316 cluster network throughput. There is little difference in CPU utilization.

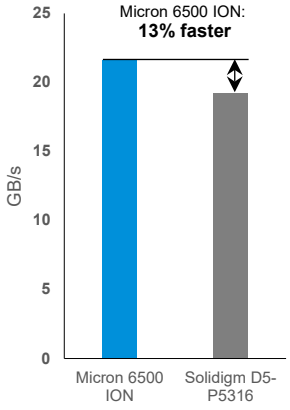

Figure 12a: Cluster throughput during rebuild

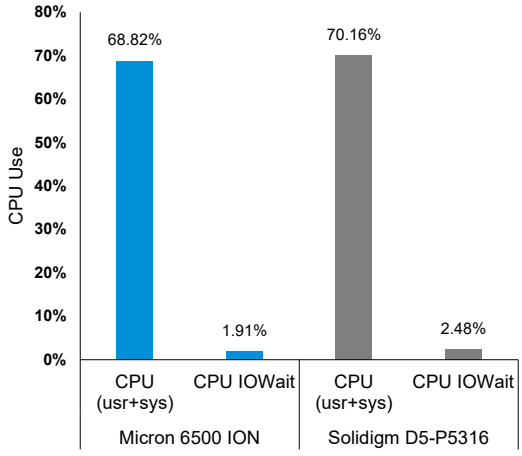

Figure 12b: CPU utilzation during rebuild

Conclusion

As software-defined storage solutions embrace the unique features and performance of SSDs, all-flash object storage solutions such as Ceph Storage Community edition enable higher performance for data analytics, artificial intelligence data lakes, and a broad variety of other workloads.

SSD-based clusters are the obvious choice when cluster performance matters, but which SSD should the cluster be built with? The answer is clear: the Micron 6500 ION.

When we compared the Micron 6500 ION to the Solidigm D5-P5316 using Ceph Storage Community Edition, we found three key areas in which the Micron 6500 ION configuration surpassed the Solidigm D5-P5316 configuration:

1. **Cluster performance:** The Micron 6500 ION configuration performance surpassed the Solidigm D5-P5316 in every tested workload. The Micron 6500 ION configuration's performance improvement ranged from a 50% higher peak for random IO up to 3.5X higher peak for sequential IO.

2. **Better CPU utilization**: The host CPUs were kept busier (higher %CPU utilization and lower %CPU IOWait) with the Micron 6500 ION cluster. Since CPUs can be a significant portion of the overall cluster hardware cost, keeping them busier and having them spend less time waiting on storage IO is beneficial.

3. **Better resiliency:** When failures occur, it is imperative to rebuild as fast as possible because extended down time increases the risk of data loss. This potential risk is mitigated by restoring cluster health as quickly as possible. The Micron 6500 ION configuration restored Ceph Storage Community Edition cluster health 5 hours faster than the Solidigm D5-P5316 configuration.

Simply put: The Micron 6500 ION cluster is faster, takes greater advantage of CPU resources, and recovers from failure quicker than the Solidigm D5-P5316 cluster. This document clearly illustrates how Micron 6500 ION SSD enables performant, fault-tolerant Ceph clusters that offer exemplary performance and resiliency.

How We Tested

Object testing utilizes the RADOS bench benchmarking tool, which is provided as part of the Ceph package, to measure object IO performance. This benchmark reports throughput performance in GB/s. Object IO uses a RADOS gateway service operating on each load generation server.

To measure object write throughput performance, each test executes RADOS bench with a "threads" value of 16 on a load generation server writing directly to a Ceph storage pool using 256KB objects. The number of RADOS bench instances is scaled from 10 to 60 to determine the maximum throughput value. Objects are purged from the pool between each test.

All other tests use 60 RADOS bench instances and execute 256KB object workloads against the storage pool while scaling RADOS bench client thread count between 2 threads and 32 threads in base-2 increments.

Test iterations execute for 10 minutes each. Before each iteration, all Linux filesystem caches are cleared. The reported results are the mathematical mean across all test runs.

Why we use erasure coding for data protection in this test

Ceph data protection focuses on continuous operation after individual data node failure (failures to tolerate [FTT]). Ceph supports two modes of data protection, including erasure coding and replication.

- **Erasure coding (EC):** EC stores data differently than replication. EC breaks an object into data chunks and coding chunks, which are then stored on different physical storage devices. If a failure occurs, the EC algorithm can use the surviving chunks to recreate the missing information. EC works well with NVMe SSDs like the Micron 6500 ION.
- **Replication**: Replication can be configured to support multiple node failures by adjusting the number of data replicas (copies of original data) to store within the Ceph cluster. The default for Ceph is 3X replication (i.e., 3 copies of all data).

We chose erasure coding because when compared to 3X replication, 4+2 erasure coding offers:

- **2X usable capacity**: This helps reduce the number of servers needed by half.[2]
- **Same level of data protection**: Erasure coding supports the same number of failures to tolerate as 3X replication.

Ceph 4+2 erasure coding details: Write operations within Ceph pools always take place on a "primary" OSD for a given client session for each of these data protection mechanisms.

Once the data is written to the primary OSD, the configured data protection algorithm is executed, and the data is distributed to the other OSD nodes for that storage pool. Ceph intelligently distributes client connections throughout the OSD nodes within the pool to help ensure that no single OSD node is overloaded.

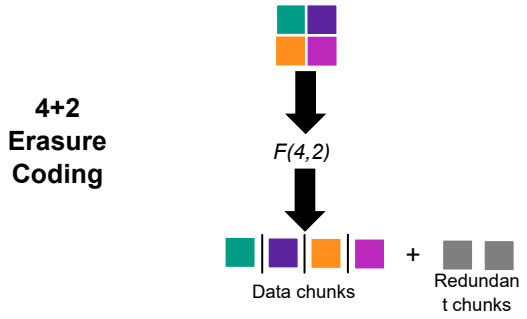

Figure 13: Ceph erasure coding example

Each data object is subdivided into four equal-sized data chunks and distributed to four different OSD nodes.

Two extra chunks, called redundant chunks, are generated using an erasure coding process and written to two different OSD nodes.

Using 4+2 erasure coding, each 1TB of usable storage capacity requires 1.5TB of raw capacity.

Test cluster configuration

The test is conducted on a Ceph Storage Community Edition cluster consisting of six data nodes that host Ceph OSDs and three monitor nodes, as illustrated in Figure 14. Load generation is created using six servers (not shown).

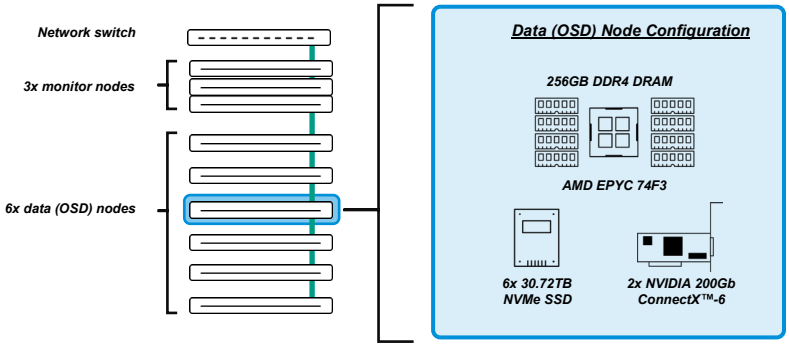

Figure 14: Ceph cluster overview

Server configuration

Table 1 describes the hardware and software configuration for each of the server types used in the test configuration. The test environment consists of six OSD (data) nodes, three monitor nodes, and six load-generation servers.

	Data (OSD) Nodes	Monitor Nodes	Load-Generation Servers
CPU Architecture	AMD EPYC® 74F3 (24-cores) Single Socket, NUMA per socket: 1 SMT: enabled IOMMU: enabled	AMD EPYC 74F3 (24-cores) Single Socket	AMD EPYC 74F3 (24-cores) Single Socket
CPU Cores per Server	24	24	24
Memory	Micron 256GB DDR4 DRAM	Micron 256GB DDR4 DRAM	Micron 256GB DDR4 DRAM
Network	2x NVIDIA® 200Gb ConnectX™-6 (MCX623105AN-VDAT)	1x NVIDIA 200Gb ConnectX-6 (MCX623105AN-VDAT)	1x NVIDIA 200Gb ConnectX-6 (MCX623105AN-VDAT)
Operating System	Ubuntu 20.04 HWE (Kernel 5.15)	Ubuntu 20.04 HWE (Kernel 5.15)	Ubuntu 20.04 (Kernel 5.15)
Boot Device	Micron 7300 PRO NVMe SSD (960GB)	Micron 7300 PRO NVMe SSD (960GB)	Micron 7300 PRO NVMe SSD (960GB)
Data Storage	6x Micron 6500 ION SSD (30.72TB) 6x Solidigm D5-P5316 SSD (30.72TB)	NA	NA

Table 1: Server and software configurations

Ceph configuration parameters

Four OSDs per SSD are configured, totaling 24 OSDs per server and 144 OSDs for the entire storage cluster. Each OSD storage node is configured as a failure domain within the Ceph infrastructure to ensure that data chunks from a protected object are stored on different server nodes. Raw storage for the Ceph cluster is approximately 1000TB. Storage pools are configured for 4+2 Erasure Coding, as shown in Table 2.

Pool Data Protection Type	Placement Groups	Usable Capacity
4+2 Erasure Coding	2,048	667TB

Table 2: Storage pool configuration

Network configuration

A single NVIDIA SN4700 400 GbE switch is used for test purposes only. It is recommended that at least two switches are used for production environments. The second switch is commonly used in production deployments only to help ensure overall cluster reliability.

micron.com/6500ION